
Proceedings Paper:
Hughes, Vincent orcid.org/0000-0002-4660-979X, Clermont, Frantz and Harrison, Philip orcid.org/0000-0003-2419-2388 (2020) Correlating cepstra with formant frequencies: : implications for phonetically-informed forensic voice comparison. In: Proceedings of Interspeech 2020. , pp. 1858-1862.

https://doi.org/10.21437/Interspeech.2020-2216

Correlating cepstra with formant frequencies: implications for phonetically-informed forensic voice comparison

Vincent Hughes[1], *Frantz Clermont*[2], *Philip Harrison*[1]

[1]Department of Language and Linguistic Science, University of York, UK
[2]School of Culture, History and Language, Australian National University, Australia
vincent.hughes@york.ac.uk, dr.fclermont@gmail.com, philip.harrison@york.ac.uk

Abstract

A significant question for forensic voice comparison, and for speaker recognition more generally, is the extent to which different input features capture complementary speaker-specific information. Understanding complementarity allows us to make predictions about how combining methods using different features may produce better overall performance. In forensic contexts, it is also important to be able to explain to courts what information the underlying features are actually capturing. This paper addresses these issues by examining the extent to which MFCCs and LPCCs can predict F0, F1, F2, and F3 values using data extracted from the midpoint of the vocalic portion of the hesitation marker *um* for 89 speakers of standard southern British English. By-speaker correlations were calculated using multiple linear regression and performance was assessed using mean rho (ρ) values. Results show that the first two formants were more accurately predicted than F3 or F0. LPCCs consistently produced stronger correlations with the linguistic features than MFCCs, while increasing cepstral order up to 16 also increased the strength of the correlations. There was, however, considerable variability across speakers in terms of the accuracy of the predictions. We discuss the implications of these findings for forensic voice comparison.

Index Terms: cepstral-coefficients, formants frequencies, speaker recognition, speaker characterisation, forensic voice comparison

1. Introduction

1.1. Complementarity of features

An important element in the development of any approach to speaker recognition is the choice of input features. For the past two or more decades, cepstral-coefficients (CCs) have been the industry standard within automatic speaker recognition (ASR) systems. Meanwhile, linguistic approaches to speaker recognition have typically focused on the componential analysis of a range of features, such as vowel formant frequencies and fundamental frequency (F0). There is now a growing trend towards the integration of ASR and linguistic approaches, with the ultimate aim of improving overall performance and some previous research has had success in this regard [1,2,3]. Implicit within such work is the question of whether different features capture complementary speaker-specific information.

Some relationships between features are predictable. CCs indirectly capture spectral information relating to the size and configuration of the supralaryngeal vocal tract. The smoothing involved in deriving CCs is claimed to decouple source from filter [4]. However, the extent of this decoupling is, in principle, determined by cepstral order, such that the more CCs extracted, the more harmonic information is modelled. Different types of CCs also provide different levels of spectral resolution. MFCCs capture more detail at the lower end of the frequency scale and are more sparse in higher frequencies. LPCCs, however, model the frequency scale in a linear way. As with CCs, formants, too, are related to the supralaryngeal vocal tract, but only capture partial information (relating to the peaks) about the entire spectrum. In principle, CCs should, in some way, also encode formant frequency information. This is consistent with the findings of [5], in which only marginal improvements in system performance were reported when fusing MFCCs with long-term formant distributions. Similarly, empirical data are consistent with the theoretical decoupling of source and filter in CCs. [6] reports potentially large improvements in system performance when combining MFCC-based ASR systems with laryngeal voice quality features.

1.2. Forensic considerations

The issue of the complementarity of features is more critical in the forensic context for two reasons. Firstly, it is essential that an expert's conclusion is an accurate reflection of the strength of the voice evidence. If multiple correlated features are analysed, there is the potential for overstating evidential value. Secondly, an expert's evidence must be understandable to a court, in order for them to make informed decisions about the *ultimate issue* of innocence or guilt. A key benefit of the linguistic approach to forensic voice comparison is that features are well-understood in terms of their mapping onto physical and anatomical properties of speech production. While CCs are known to be a representation of the spectrum, they are abstract mathematical values derived through various levels of processing (e.g. iFFT). As such, CCs do not map in any straightforward way onto physiological properties of speakers or the articulatory implementation of speech production, making ASR evidence more difficult to explain to courts.

1.3. This study

Previous work has attempted to directly predict formants and F0 from CCs. Performance has generally been assessed using the correlation between measured and predicted values for the linguistic features. Research has shown that correlations are strongest when using phoneme-specific and speaker-dependent modelling, with [7,8] reporting correlation coefficients of over 0.9. Across studies, F1 and F2 are predicted more accurately than F3, while F0 produces the weakest correlations [9]. This is consistent with the theoretical decoupling of source and filter in

deriving CCs. However, the fact that F0 is at all predictable indicates that some source information is captured.

The present study continues this line of enquiry to better understand the relationship between CCs and linguistic features, but expands on previous work in a number of key ways: (i) We use spontaneous, more forensically realistic speech material rather than controlled, lab speech; (ii) We analyse a vowel segment (the hesitation marker *um*) that has considerable speaker-discriminatory power and so is useful in forensic voice comparison; (iii) We directly compare the predictive power of LPCCs and MFCCs, using different cepstral orders. This is important because not all ASR systems use the same underlying features or the same number of CCs. Further, there are theoretical predictions associated with different types and orders of CCs, as outlined above; (iv) Finally, we also examine the performance of individual speakers and consider the implications of our results for forensic voice comparison.

2. Method

2.1. Database

A total of 89 young, male speakers of Standard Southern British English (modern Received Pronunciation) from the Dynamic Variability in Speech (DyViS) database [10] were used (11 speakers from the full 100 available were not included due to insufficient numbers of tokens). The database was collected for forensic phonetic research and speakers engaged in forensically realistic tasks: a mock police interview (Task 1) and a telephone conversation with an accomplice (Task 2). Both tasks involved spontaneous, conversational speech of between 9 and 30 minutes in duration. For both tasks, we used high-quality, studio recordings (44.1kHz sampling rate, 16-bit depth) to remove confounding effects related to measurement error with poorer quality materials.

2.2. Hesitation markers

The hesitation marker *um* was analysed, principally because it has been shown to carry considerable speaker-specific information. Using good quality materials, [12] report equal error rates of as low as 4.08% and log LR cost (C_{llr}) values of as low as 0.12 using the acoustics of the vocalic portion of *um* alone (*um*, with the nasal /m/ following the vocalic portion, was found to perform better than *uh*, which is entirely vocalic). When fused with an MFCC-based ASR system, segmental analysis of *um* has also been shown to improve performance compared with the ASR system in isolation [1]. Alongside this, there are a number of reasons why hesitations are useful for the purposes of forensic voice comparison. Firstly, these hesitation phenomena occur frequently (around 3.7 occurrences per minute; [11]). Secondly, they are thought to be produced below the level of consciousness and so are relatively resistant to disguise. Thirdly, they often occur adjacent to silences, making their formants easy to measure and less susceptible to coarticulatory effects. This, in turn, helps to reduce the amount of within-speaker variability that they exhibit.

2.3. Feature extraction

A total of 6758 *um* tokens (median N tokens across recordings of both tasks = 70 per speaker, max = 159, min = 26) were analysed by manually marking the onset and offset of the vocalic portion. The first three formants were then extracted from a 20ms frame at the temporal midpoint of the vowel. Values were extracted in *Praat* [13] using the *Formant: Burg* function with an LPC order of between 10 and 12, determined on a speaker-by-speaker basis to ensure measurements were as reliable as possible. F0 was extracted within a frequency range of 75-200 Hz using the STRAIGHT algorithm [14] in VoiceSauce [15]. From the same 20ms midpoint frame, MFCC and LPCC vectors up to order 16 were extracted in MATLAB. This involved downsampling the recordings from 44.1 kHz to 8 kHz, such that CC extraction was performed within a 0-4000 Hz range.

2.4. Cepstrum-to-F{0,1,2,3} mapping

Cepstrum-to-F{0,1,2,3} mapping was performed by pooling the Task 1 and Task 2 data for each speaker. Although the two tasks were recorded in separate sessions, the channel characteristics are essentially identical and pooling data allowed us to maximise the number of tokens available. The MFCC and LPCC vectors were used to predict the univariate F{0,1,2,3} values by-speaker, using a multiple linear regression model based on the weighted sum of the CCs. The formulation of the regression model is:

$$F(i) = a_{i,0} + a_{i,1}\mathrm{x}(1) + \ \ldots + a_{i,M}\mathrm{x}(M) + \varepsilon \quad (1)$$

where $F(i)$ is the dependent variable, i.e. F{0,1,2,3}, $[a_{i,0}, \ldots, a_{i,M}]$ are regression coefficients for the feature i, x is the vector of CCs of length M and ε is an error term. The regression coefficients are determined using least squares estimation. The regression model can then be used to predict values for $F(i)$. The correlation between the predicted values and the measured values was obtained and represented as a correlation coefficient (ρ), whereby the closer the value to 1 the better the predictive power of the model.

Regression models were trained and tested on the same data for two reasons. Firstly, it allowed us to maximise the amount of data available. Secondly, our aim was to understand and explore the relationships between the underlying features, not to build a system with the best predictive performance. We also used a speaker-dependent method, as this has been shown to generate stronger correlations than speaker-independent methods [7,8]. The overall strength of the cepstrum-to-F{0,1,2,3} mapping was measured using mean ρ across all speakers. This process was repeated for each linguistic feature (F{0,1,2,3}) using MFCCs and LPCCs of different orders.

3. Results

3.1. MFCCs vs. LPCCs

Table 1 gives a summary of the mean ρ values across all 89 speakers for each of the linguistic features, using MFCCs and LPCCs as inputs to the best model with all 16 coefficients.

Table 1: *Mean ρ for LPCC-to-F{0,1,2,3} and MFCC-to-F{0,1,2,3} (both using 16 coefficients)*

	F1	**F2**	**F3**	**F0**
LPCC	0.909	0.925	0.865	0.803
MFCC	0.891	0.903	0.838	0.829

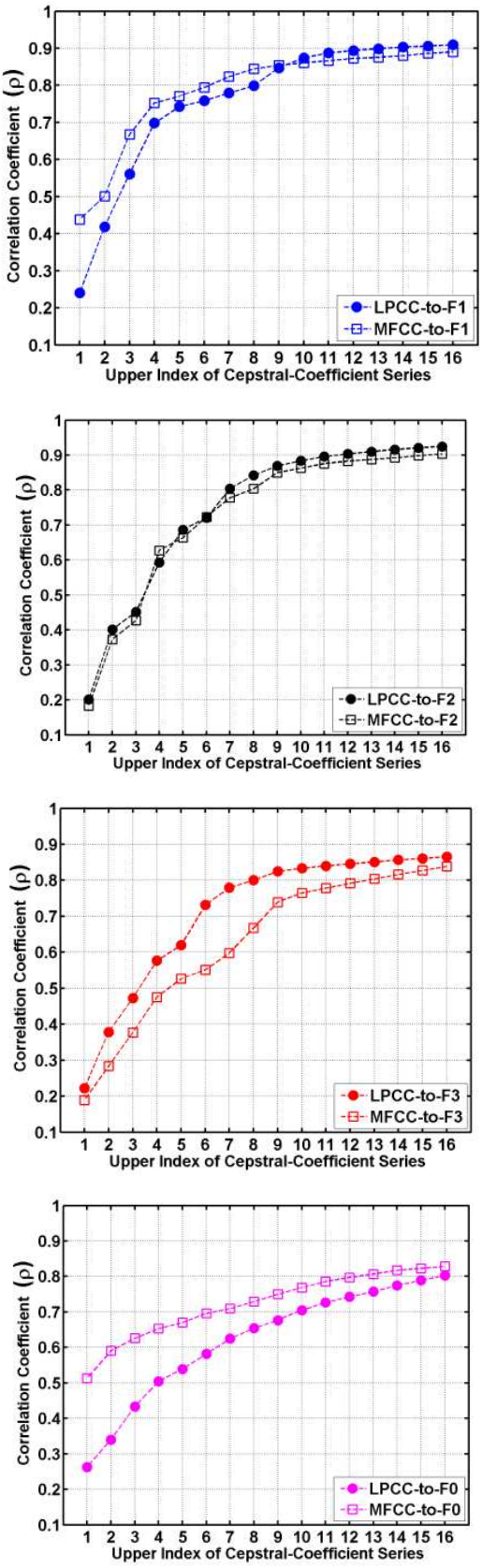

Figure 1: *Mean ρ across the 89 speakers for F{0,1,2,3} using LPCCs and MFCCs as a function of the upper index of the cepstral-coefficient series.*

For all three formants, the LPCCs provide marginally better predictive performance than the MFCCs, while the MFCCs perform slightly better than the LPCCs for F0. The mean ρ values squared afford the further interpretation that the models explain 75% to 86% of the variability in the formants using the LPCCs, and 70% to 82% using the MFCCs. For F0, the models account for 69% of the variability using the MFCCs and 64% using the LPCCs.

3.2. Effects of cepstral-coefficient order

Figure 1 shows the effect of increasing cepstral order on the mean ρ values for both LPCCs and MFCCs. There is continual improvement in the correlation between predicted and measured F{0,1,2,3} as the number of CCs increases. After around 10 CCs, the rate of improvement begins to decrease, although the ρ values when using 16 CCs are still the highest. The MFCCs provide better performance when using smaller numbers of CCs, compared with LPCCs at least for F0 and F1.

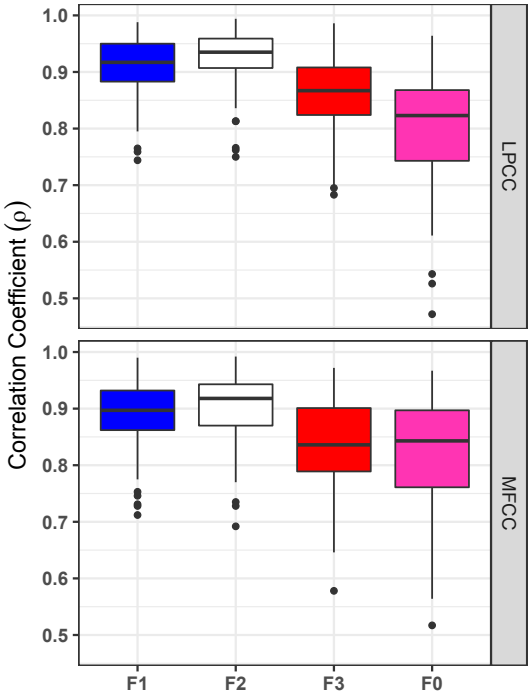

Figure 2: *Distributions of ρ values for the 89 speakers using LPCCs (above) and MFCCs (below) (based on vectors of 16 CCs) to predict F{0,1,2,3}.*

3.3. Individual features

The best performance is achieved for F2, followed by F1, F3 and finally F0. One explanation for why F1 and F2 outperform F3 in our data may be due to the accuracy of the original formant measurements. Formants are generally more problematic to estimate in the higher frequencies, as they often have lower amplitude and wider bandwidths. Further, the formant measurements here were automatically extracted using fixed settings by-speaker (i.e. settings differed from speaker to speaker). Thus, there was no hand correction of the formant data which would, undoubtedly, have improved the accuracy of the measurements [16]. For F2, the correlations for the MFCCs and

LPCCs are essentially the same across cepstral orders, while for F3 the LPCCs consistently produce the strongest correlations.

3.4. Individual speakers

Figure 2 shows the distributions of ρ values across speakers using both LPCCs and MFCCs (using 16 CCs) to predict F{0,1,2,3}. Notably, there is considerable variation, with some speakers producing near perfect correlations between the predicted and measured F{0,1,2,3} values, and other producing much weaker correlations. In line with the findings in §3.3, the greatest variability is found for F0 and F3. Much narrower ranges of variability are found for F1 and F2. This pattern is consistent across both LPCCs and MFCCs.

4. Discussion

The results of this study have shown that F{0,1,2,3} can be predicted from vectors of LPCCs and MFCCs with a relatively high degree of accuracy. The mean ρ values in Table 1 compare well with the speaker-dependent correlations reported in [7,8], with ρ values of over 0.9 in some cases (meaning that the MFCCs and LPCCs were able to explain over 80% of the variability in some of the formant data. This performance is extremely impressive given the large number of tokens, the use of spontaneous speech, the degree of between-speaker variability displayed by the hesitation markers and the fact that formant data were automatically extracted.

A number of general patterns were also found in our results. LPCCs marginally outperformed MFCCs when predicting the formants (although MFCCs performed best for F0). This is likely due to the fact that the LPCCs are based on the linear prediction (LP) model, which is particularly good at representing spectral peaks, due to the all-pole constraint. The hesitations markers examined here are, in many ways, an ideal case for the LP model, since they tend to display widely spaced, and therefore easily identifiable, formants (means across all speakers: F1 = 608Hz, F2 = 1378Hz, F3 = 2496Hz). The results in §3.2 reveal an interaction between the predictive strength of the input and cepstral order. For F1, higher correlations were found for the MFCCs when the upper index of the CC series was low, whereas the LPCCs performed better with larger numbers of CCs. This shows that while LPCCs may generally perform better at predicting formants, greater spectral resolution is needed to approach (and ultimately outperform) the MFCCs. No such interaction was found for F2 or F3. In terms of the individual linguistic features, F2 produced the largest ρ values irrespective of the input, followed by F1, then F3, and finally F0. This ordering is also consistent with [7,8].

These findings validate previously-reported findings that the linearity of cepstrum-to-F{0,1,2,3} mapping is more consistent within speakers and stronger within a reduced phonetic space, such as that which is spanned by *um*. The relationship with formants is expected given that they, along with CCs, in theory capture information about the supralaryngeal vocal tract. The finding that F0 can also be predicted from CCs suggests that, in practice, the decoupling of source and filter in deriving CCs is not absolute. The increase in the strength of the correlation for F0 as a function of cepstral order is consistent with the assertion in §1.1, that the degree of smoothing involved in deriving CCs affects the extent to which source and filter can be decoupled. Higher orders of CCs provide more detailed spectral resolution that also models some harmonic structure.

An interesting finding, that has not been addressed in previous work, is that there is considerable between-speaker variability in the predictive power of the cepstrum-to-F{0,1,2,3} mapping. This variability appears to be dependent on the linguistic feature being predicted. That is, speakers who produce large ρ values for F0 do not necessarily produce large ρ values for formants. The same is true of the individual formants. As outlined above, one key factor determining the success of the cepstrum-to-F{0,1,2,3} mapping is the accuracy of the raw F{0,1,2,3} data. It is well known that formants are better tracked for some speakers than others. Using the same recordings as the present study, [17] showed that this proclivity towards formant measurement errors due to the settings used can have dramatic effects on a speaker's performance within a formant-based speaker recognition system (and on the overall performance of the system). There is some overlap between the problematic speakers in [17] and the speakers who generally produce the weakest correlations in the present study. This highlights the importance of accurate measurement both in terms of providing reliable forensic evidence, but also for understanding what information our systems are actually capturing.

5. Conclusions

This study has demonstrated the existence of linear relationships between the cepstrum and each of the formants, using a large amount of automatically extracted data from a forensically valuable segment (u*m*) and forensically realistic speaking tasks. A similar but weaker trend can be said about F0. The strong correlations resulting from the linear mappings have relevant implications for forensic voice comparison.

First, they argue against the value of the formants in favour of the cepstrum in ASR. The lack of complementarity in our study is consistent with the findings of [5] that show no additional benefit of fusing formants may be expected with CC-based ASR systems. Thus, we conclude that the improvements gained by fusing linguistic features with an ASR system in [1,2,3] are principally due to the segmental nature of the linguistic analysis, compared with the holistic approach employed by the ASR system, rather than the fundamental complementarity of the features. Perhaps more positively, the relationships observed here confirm that the cepstrum does encode the bulk of phonetic and articulatory information carried by the formants (as well as showing extremely good speaker discriminatory power; the value of segemental cepstra is shown in [18]). Forensic evidence based on cepstrum-based systems is therefore amenable, albeit indirectly, to phonetic and articulatory interpretations. Last but not least, the consistency of a linear cepstrum-to-formant mapping within speakers raises the possibility that the forensic practitioner might be able to estimate or to validate the formants for a speaker's new segments from an existing predictive model for that speaker. Further work will be necessary to investigate this practical benefit.

6. References

[1] V. Hughes, S. Wood, and P. Foulkes, "Formant dynamics and durations of um improve the performance of automatic speaker recognition systems," in *Proceedings of the Australasian Conference on Speech Science and Technology, 7-9 December, Sydney, Australia*, 2016, pp. 249–252.

[2] J. González-Rodriguez, "Speaker recognition using temporal contours in linguistic units: the case of formant and formant-

bandwidth trajectories," in *INTERSPEECH 2011*, *12th Annual Conference of the International Speech Communication Association, August 27-31, Florence, Italy,* 2011, pp. 133–136.

[3] C. Zhang, G. S. Morrison, and T. Thiruvaran, "Forensic voice comparison using Chinese /iau/," in *Proceedings of the International Congress of Phonetic Sciences, August 17-21, Hong Kong*, 2011, pp. 2280–2283.

[4] D. Jurafsky and J. H. Martin, *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition* (2nd ed). New Jersey: Prentice-Hall, 2009.

[5] V. Hughes, P. Harrison, P. Foulkes, J. P. French, C. Kavanagh, and E. San Segundo, "Mapping across feature spaces in forensic voice comparison: the contribution of auditory-based voice quality to (semi-)automatic system testing," in *INTERSPEECH 2017*, *18th Annual Conference of the International Speech Communication Association, August 20-24, Stockholm, Sweden,* 2017, pp. 3892–3896.

[6] V. Hughes, A. Cardoso, P. Foulkes, J. P. French, P. Harrison, and A. Gully, "Forensic voice comparison using long-term acoustic measures of voice quality," in *Proceedings of the International Congress of Phonetic Sciences, August 4-10, Melbourne, Australia,* 2019, pp. 1455–1459.

[7] J. Darch and B. Milner, "Analysis and prediction of acoustic speech features from mel-frequency cepstral coefficients in distributed speech recognition architectures," *Journal of the Acoustical Society of America*, vol. 124, no. 6, pp. 3989–4000, 2008.

[8] F. Clermont, "Cepstrum-to-formant mapping of spoken vowels," paper presented at the Conference of the International Association in *IAFPA 2013 – 22nd Annual Conference of the International Association for Forensic Phonetics and Acoustics, July 21-24, Tampa, Florida,* 2013.

[9] D. J. Broad and F. Clermont, "Formant estimation by linear transformation of the LPC cepstrum," *Journal of the Acoustical Society of America*, vol. 86, no. 5, pp. 2013–2017, 1989.

[10] F. Nolan, K. McDougall, G. de Jong, and T. Hudson, "The DyViS database: style-controlled recordings of 100 homogeneous speakers for forensic phonetic research," *International Journal of Speech, Language and the Law*, vol. 16, no. 1, pp. 31–57, 2009.

[11] N. Tschäpe, J. Trouvain, D. Bauer, and M. Jessen, "Idiosyncratic patterns of filled pauses," in *IAFPA 2005 – 14th Annual Conference of the International Association for Forensic Phonetics and Acoustics, August 3-6, Marrakesh, Morocco,* 2005.

[12] V. Hughes, S. Wood, and P. Foulkes, "Strength of forensic voice comparison evidence from the acoustics of filled pauses," *International Journal of Speech, Language and the Law*, vol. 23, no. 1, pp. 99–132, 2016.

[13] P. Boersma and D. Weenink, Praat: doing phonetics by computer [computer program], version 6.0.49, 2019.

[14] H. Kawahara, A. de Cheveigne, and R. D. Patterson, "An instantaneous-frequency-based pitch extraction method for high-quality speech transformation: revised TEMPO in the STRAIGHT-suite," in *Proceedings of the International Conference on Spoken Languages Processing, November 30 – December 4*, *Sydney Australia,* 1998.

[15] Y. Shue, *The Voice Source in Speech Production: Data, Analysis and Models*, PhD thesis, University of California, Los Angeles, 2010.

[16] P. Foulkes, G. Docherty, S. Shattuck-Hufnagel, and V. Hughes, "Three steps forward for predictability: consideration of methodological robustness, indexical and prosodic factors, and replication in the laboratory," *Linguistics Vanguard*, vol. 4, no. 2, 2018.

[17] V. Hughes, P. Harrison, P. Foulkes, J. P. French, and A. Gully, "Effects of formant analysis settings and channel mismatch on semi-automatic forensic voice comparison," in *Proceedings of the International Congress of Phonetic Sciences, August 4-10, Melbourne, Australia,* 2019, pp. 3080–3084

[18] P. Rose, "More is better: likelihood ratio-based forensic voice comparison with vocalic segmental cepstra frontends," *International Journal of Speech, Language and the Law,* vol. 20, no. 1, pp. 77-116, 2013.